Timed Probabilistic Automaton: a Bridge between Raven and Song Scope for Automatic Species Recognition

Shufei Duan, Jinglan Zhang, Paul Roe, Jason Wimmer, Xueyan Dong,
Anthony Truskinger, Michael Towsey
Microsoft QUT eResearch Center, Queensland University of Technology
shufei.duan@student.qut.edu.au

Abstract

Raven and Song Scope are two automated sound analysis tools based on machine learning technique for environmental monitoring. Many research works have been conducted upon them, however, no or rare exploration mentions about the performance and comparison between them. This paper investigates the comparisons from six aspects: theory, software interface, ease of use, detection targets, detection accuracy, and potential application. Through deep exploration one critical gap is identified that there is a lack of approach to detect both syllables and call structures, since Raven only aims to detect syllables while Song Scope targets call structures. Therefore, a Timed Probabilistic Automata (TPA) system is proposed which separates syllables first and clusters them into complex structures after.

1 Introduction

Animal call recognition plays a significant role in environmental monitoring where it can be used as an indicator of species diversity, abundance and overall environmental health (Towsey et al., 2012). Manual analysis is effective for single species identification but failed to deal with datasets over large spatiotemporal scale. Automatic tools greatly facilitate animal call recognition especially over large datasets by reducing the process time and increasing the efficiency.

Two state-of-the-art developments: Raven (Bioacoustics Research Program, 2011) and Song Scope (Wildlife Acoustics, 2011) are developed for assisting ecologists' work in dealing with a large amount of data and average people conducting their research on animal sounds analysis. However, though they have been widely used for years, no actual case study has been done for comparing their performance on a real world data (Crothers, Gering, & Cummings, 2011; Depraetere et al., 2012). The aim of this paper is to explore the performance of these tools and potential application areas using a real world dataset. Through exploration we found that Raven and Song Scope are especially built for either syllable or call structure, not for both. To build a bridge between them, we present Timed Probabilistic Automata (TPA) to join syllable and call structure detection together.

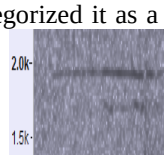
2 Call Structures

Many animal calls have hierarchical structures. A typical bird song is divided into phrases, syllables, and elements (Somervuo, Harma, & Fagerlund, 2006). Generally, syllables mean timestamps in an audio stream (Zhuang et al., 2010) while call structures consist of single or multiple syllables.

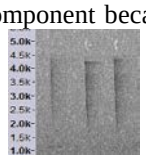
Since animal call structures comprise of some common patterns, there are many attempts to define these typical components (McCallum, 2010; Scott Brandes, 2008). Different from definitions in the aspect of phonetics, Duan defined broad acoustic components according to their appearance in the spectrogram (Duan et al., 2011). These components can be divided into two parts. The first part called primitive components include whistle (a horizontal line), click (a vertical line), slur (from the whip to a slow chirp), warble (moduled in one direction and then back again), and blocks (energy concentrated rectangular or triangular areas) while composite ones include stacked harmonics (vertical stacks of lines or warbles spaced equally) and oscillations (horizontal repeated acoustic components). Specifically, Figure 1 gives out the appearance of components and the typical species whose call structures are comprised of these components.

In fact, primitive components as well as stacked harmonics are similar to the common definition of syllable. They are inseparable in time and can be used to construct call structures. Oscillation is a special component which actually a quite common call structure

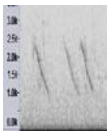
among animal calls. It consists of primitive components typically clicks or stacked harmonics. Duan categorized it as a component because to detect this pattern is also fundamental in animal call recognition. The name of these components follows McCallum's definition.

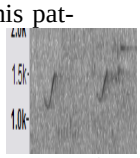


EasternWhipbird, Little Bronza-Cuckoo, Gray Fantail.
(a)Whistle

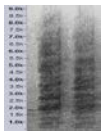


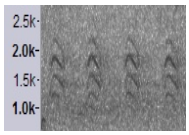
Eastern Yellow Robin, Eastern Whipbird, Mistletoebird.
(b) Click


Little Friarbird, White-browed Scrubwren.
(c) Slurs

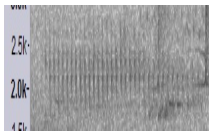


Indian Peafowl, Eastern Koel, Australian Owlet-Nightjar.
(d) Warble


Rainbow lorikeet, Masked Owl.
(e) Block



Australian Owlet-Nightjar, Corellla sp, Crow.
(f) Stacked Harmonic



Lewin's HoneyEater, Rufous Whistler, White-browed Scrubwren, Little Friarbird.
(g) Oscillation

Figure 1. Acoustic Components' Appearance and Representative Species.

3 Software

Raven and Song Scope are the state-of-the art tools for terrestrial animal bioacoustics analysis, which are worth comparing their performance so as to provide scientific foundations for the future research.

3.1 Raven

Raven, produced by the Cornell Lab of Ornithology, is a software program for the acquisition, visualization, measurement, and analysis of sounds (Charif et al., 2010). Raven centers around audio files viewed as waveforms and spectrograms, and allows users to apply a set of analysis tools. It is designed for birdsong analysis workflows, so for example it provides tools to perform band-pass filters and manual or semi-automatic syllable segmentation (Stowell & Plumbley, 2011). Raven has very friendly interface and easy to learn for use. Besides, it has very powerful play and cut modules for users to focus on the specific fraction which is mostly amenable to analysis. In terms of target detection, Raven has two detectors: a band limited energy and an amplitude detector.

- Band Limited Energy Detector
  It estimates the background noise of a signal and uses this to find sections of signal that exceed a user-specified signal to noise ratio threshold in a specific frequency band, during a specific time.
- The Amplitude Detector
  It detects regions of a signal where the magnitude of the waveform's envelope exceeds a threshold value.

These detectors are relatively simple to set up as long as following the manual introduction. The process to run the band limited energy detector is based on the spectrogram while the amplitude one works on the waveform.

Another important function of Raven is to conduct the batch processing which allows users to run the detector over a large scale of dataset, and this benefits the analysis work on a large amount of data.

Raven aims to detect syllables. Multiple detectors can be run over one spectrogram (waveform), which allows one to build separate detectors for different syllables. However, it is limited for detecting call structures that contain multiple syllables. Even if different syllables are picked out, they are not jointed together to form a call structure.

3.2 Song Scope

Song Scope, produced by Wildlife Acoustics, Inc, is a sophisticated digital signal processing application designed to quickly and easily scan long audio recordings made in the field and automatically locate vocalizations made by specific bird species and other wildlife (Song Scope 4.0 User's Manual, 2011). Compared with Raven, Song Scope does not have general purpose recording or play back controls. What is more, it does not allow users to replay particular sections without annotating these sections and saving them as new files. Song Scope also centers on audio files viewed as waveforms and spectrograms. The interface is simple and colorful. However, the colorful spectrogram is relatively not comfortable for visualization compared with a gray-scale alternative.

Song Scope is aiming for detecting call structures, which is different from Raven. The Song Scope classification algorithms are based on Hidden Markov Models using spectral feature vectors similar to Mel Frequency Cepstral Coefficients as these methods have been proven to work effectively in robust speech recognition applications (Agranat, 2009).

We noticed that Song Scope segments the syllables first and clusters those together to form call structures. However, we can easily discover that this approach is very sensitive to the purity of syllables. If syllables are polluted by non-target species or background noise, the model would be also very sensitive, thereby affecting the recognition accuracy. This is the reason why Agranat (developer from Wildlife Acoustic, Inc) chose very clean datasets for testing (Agranat, 2009). The dataset they used were cut to individual vocalizations manually and in each vocalization, there is rare other

species call. Thus, the clustering algorithm works relatively well as there is no pollution.

The process to run Song Scope is not hard but it does require users have some background knowledge of signal processing to know and set up parameters. Song Scope also supports batch processing to deal with large scale of data.

Regarding to the annotation work, both Raven and Song Scope cannot accept existing call tags. This is inconvenient to share work among different research groups. In our case, we have already collected a library of tags which were labeled by bird watchers. The quick and convenient way is that we should directly import these tags into the software and don't need to label them twice.

4 Experiments

This experiment is set up to explore the actual performance of Raven and Song Scope on real world dataset.

4.1 Dataset

The testing dataset was collected from the Samford Valley (20 kilometres north-west of Brisbane, Queensland, Australia) during the dawn chorus from 4am to 9am, 14th, Oct, 2010. This is a dataset tagged by a team of bird watching enthusiasts. About 47 species vocalized during this period. Among these species, five representative samples were selected to characterize different types of call structures as mentioned in section 2 (see Figure 1), Lewin's HoneyEater for oscillations, Eastern Whipbird for whistles and clicks, Eastern Koel for warbles, Torresian Crow for stacked harmonics, and Rainbow Lorikeet for blocks. There are in total 131 minute recordings which contain HoneyEater call, 167 minutes for Whipbird calls, 237 minutes for Koel calls, 67 minutes for Crow calls, and 93 minutes for Lorikeet calls. The training dataset were selected from the same site but from a different day. Each species has 25 samples for training. This dataset is accessible on request.

Figure 2 shows the signal to noise ratio (SNR) distribution for the five hours of dawn chorus. The x-axis represents the time range in 10 minute interval from 240th minute (4am) to 540th minute (9am). The y-axis represents the SNR (in dB). The average SNR is 13 dB, while the maximum is 33 dB and the minimum is 3.7 dB. As we can see, there are three peaks located at periods: (290, 310), (390,420), and (450, 470). The minimum value for peak time is about 23 dB. This means there are many species calling at the same time and potentially cause inaccurate detection results. In fact, to precisely detect target during dawn chorus is a very difficult task for automated tools and still an unsolved research problem in the automated species detection area.

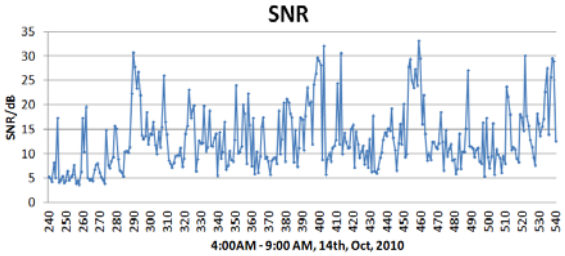


Figure 2. Noise Distribution over Dawn Chorus (4am-9am), 14th, Oct, 2010.

4.2 Software Version and Parameter Setting

The versions are: Raven Pro 1.4 and Song Scope 4.1.1, respectively. As they are parametric so setting up suitable parameters is another issue when building a recognizer. The good thing is that there is always guidance in the manual to help users set up. To configure these parameters, 25 training samples for each species were selected from the same site but not the same day (14th Oct, 2010). Table I lists the parameters we set for different species and tools.

Table I. Parameters for Different Species and Tools.

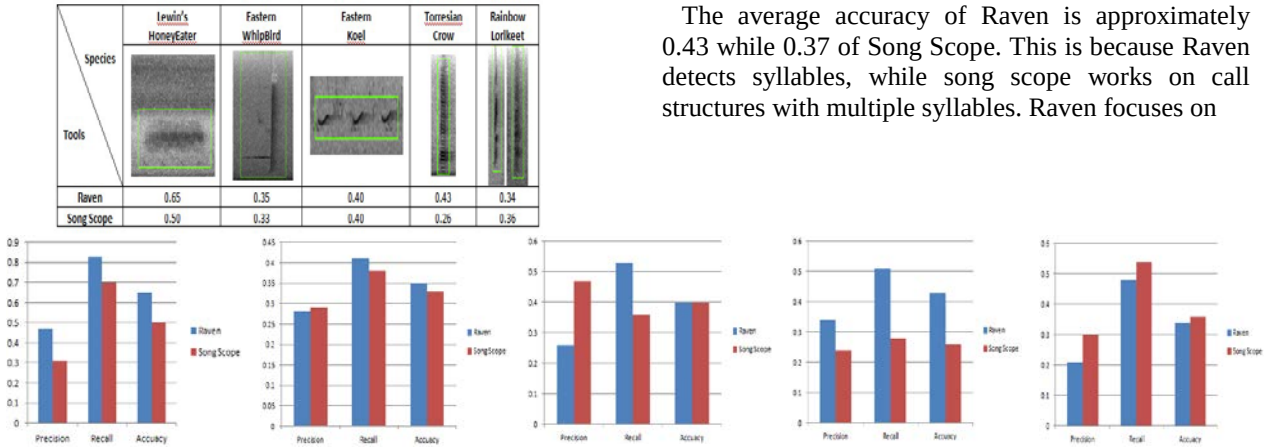
Tools	Parameters	Lewin's HoneyEater	Eastern WhipBird	Eastern Koel	Torresian Crow	Rainbow Lorikeet
Raven	Min Frequency (Hz)	1000	290	765	443	1327
	Maximum Frequency (Hz)	3100	4135	1800	3216	9125
	Minimum Duration (s)	2.57	1.0	0.25	0.19	0.20
	Maximum Duration (s)	10.82	2.5	1	0.33	0.46
	Minimum Separation (s)	0.38	0.17	0.15	0.06	0.06
	Minimum Occupancy (%)	15	20	20	20	20
	SNR threshold (dB)	5	5	5	5	5
	Block size (s)	33	10	3.5	1.0	1.5
	Hop size (s)	11	5	2.5	0.7	1.0
	Percentile	10	10	20	30	20
Song Scope	FFT size	256	1024	1024	512	512
	FFT overlap	½	½	½	½	½
	Frequency Minimum	11	20	60	12	38
	Frequency Range	24	140	37	104	167
	Amplitude Gain (dB)	0	0	0	0	0
	Background Filter (s)	1	4	1	1	1
	Max Syllable (ms)	23	2000	736	448	360
	Max Syllable Gap (ms)	12	488	10	32	46
	Max Song (ms)	2601	2067	800	448	453
	Dynamic Range (dB)	15	20	20	18	20
	Maximum Complexity	48	32	48	32	48
	Maximum Resolution	6	6	6	6	20
	Total training result	76.80±9.57%	83.32±2.27%	82.64±4.11%	81.47±3.09%	81.84±4.99%

4.3 Results

Table II gives out the accuracy of different tools for five species. Figure 3 shows the comparison of precision, recall, and accuracy for each species. The definition of precision and recall is indicated in (Olson & Delen, 2008).

According to Table II, a clear signal can be seen is that on average, the performance of Raven and Song Scope is similar across species. Raven works better than Song Scope for picking up Lewin's HoneyEater (oscillations) and Torresian Crow (stacked harmonics). Song Scope is slightly better picking up blocks (Rainbow Lorikeet). They have quite close ability to pick up Eastern Koel (warbles) and Eastern Whipbird (lines).

Table II. Accuracy of Tools for Sample Species Detection.



(A) Lewin's Honeyeater (B) Eastern Whipbird (C) Eastern Koel (D) Torresian Crow (E) Rainbow Lorikeet

Figure 3. Precision Recall and Accuracy of each Species

small sections of energy while Song scope models the structure among syllables. Noisy training samples affect the test results of Raven, but the effect is much greater on Song Scope as Hidden Markov Model views all the noisy signals as syllables and models them as a call structure. This is why Agranat chose very clear vocalizations for testing. Those vocalizations only contain target's call with no or rare pollution from other species.

When it comes to Figure 3, the average precision of Raven is approximately 0.25, which means the false positive rate is high. This is reasonable due to the feature of energy. Acoustic events exceeding the threshold will be all picked up as Raven doesn't care about the internal structures. Recall is high, which is approximately 0.50. This reflects that it can detect half the amount of the target. If users are aiming to detect the activity of targets, Raven may be more suitable.

The precision of Song Scope is around 0.32, relatively higher than Raven, which reflects low false positives rate. The recall is much lower than Raven. Low recall means the ability to detect calls is weaker than Raven. However, once a call is detected, the signal is more likely to be a true positive. If users want to detect the presence of targets, Song Scope may be suitable.

5 Timed Probabilistic Automata (TPA)

Raven target syllables, however, they cannot join these syllables together to form call structures. Song Scope detects call structures by clustering syllables, but it fails to accurately separate syllables. These tools were developed facing all types of syllables and call structures. The average performance is acceptable but unsatisfactory. The critical gap here is lack of an approach to join these two aspects together in order to have better recognition result. Timed Probabilistic Automata (TPA) are developed to solve this problem. It not only allows users to run the syllable detectors, but also give users the initiative to build call structures by themselves.

5.1 Theory and Process

TPA was adapted from the theories of Syntactic Pattern Recognition and Markov Model. Though the syntactic complexity of birdsongs cannot be directly compared with human speech due to a lack of semantics and lexicon (Berwick et al., 2011), the call structures of many avian species can be modeled by low-order Markov chains. This implies the full power of human speech recognition is probably not needed. For many instances very simple recognizers may be suitable.

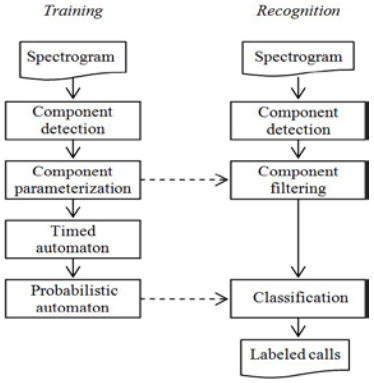


Figure 4. System Overview.

The core part of TPA is acoustic component detectors. These detectors are developed especially for five types of components: lines, warbles, blocks, oscillations, and stacked harmonics. Acoustic component detectors work as filters in the spectrogram. They are all parametric and relatively easy to configure according to the specific targets.

The processes of TPA for automatic animal call recognition are shown in Figure 4. The training and recognition part follow the same processes.

(1) According to the target's call structure shown in the spectrogram, select the proper acoustic component detector.
(2) Execute the component detector. The result of the detector is a list of components found in the spectrogram. These components are characterized by a tuple: (shape, start time, duration, minimal frequency, maximal frequency).
(3) Component filtering. Choose the training samples and train them to filter out components that do not belong to the target species.
(4) Using a timed automaton to model and control time duration of the whole target call.
(5) Apply probabilistic automata to represent the target species call structure in a sentence way.
(6) Similarity matching. Match the testing representation with the training one. If the probability falls in the training probability distribution, a target call is recognized.

5.2 Eastern Whipbird

The Eastern Whipbird is a good example for showing how to apply TPA to recognize targets. The call structure of whipbird contains a whistle and a click (see Table II). The state transition diagram is shown in Figure 5. P(w), P(g), P(c) denote the probability of whistle, gap, and click, respectively.

$$P(Whipbird) = \mathrm{P}(w) \times \mathrm{P}(g) \times P(c) \qquad (1)$$

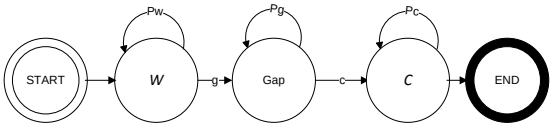


Figure 5. TPA for Eastern Whipbird.

First we call the whistle and click detectors for the component toolbox to detect whistles and clicks. Model whistle, click and gap in-between using frequency and time information which have already collected in the tuple. The TPA is applied as followed:

(1) Whistle filtering. Calculate the probability of all testing whistles. Compare this probability with the training probability. If the testing probability value is between the maximum and minimum value of training probability, a confirmed Whipbird whistle hits. Remove all irrelevant whistles.
(2) Click filtering. Calculate the probability of all testing clicks. Compare this probability with the training probability. If the testing probability value is between the maximum and minimum value of training probability, a confirmed Whipbird click hits. Remove all irrelevant clicks.
(3) Gap filtering. Calculate the probability of all gaps between Whipbird whistles and clicks. Compare this probability with training gap probability. If the testing probability value is between the maximum and minimum value of training probability, a confirmed whipbird gap hits. According to this confirmed gap value, keep pairs of whistle and click which have the confirmed gap. Remove all irrelevant whistles and clicks.
(4) Marquee the left pairs of whistles and clicks as Eastern Whipbird call.

Figure 6 gives experimental results of Eastern Whipbird recognition. Blue dots are signals left after noise removal. Green lines represent whistle and clicks. The red marquee covers Whipbird call.

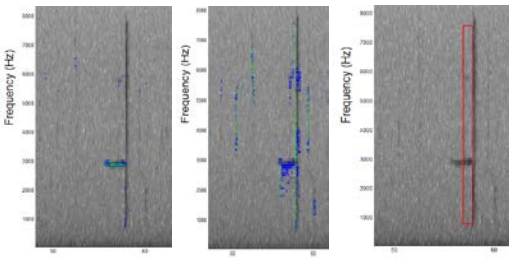


Figure 6. Eastern Whipbird Recognition by TPA

5.3 Comparison with Raven and Song Scope

To test the performance of TPA, we have compared it with Raven and Song Scope. Experiments were executed under the same conditions using the same training and testing dataset as in section 4. Table III lists the statistics of these three tools. To better illustrate points, we graph the comparison results and add the error bars with standard deviation in Figure 7.

Table III. Statistical of Tools for Eastern Whipbird

	Raven	Song Scope	TPA
Precision	0.28	0.29	0.45
Recall	0.41	0.38	0.59
Accuracy	0.35	0.33	0.52

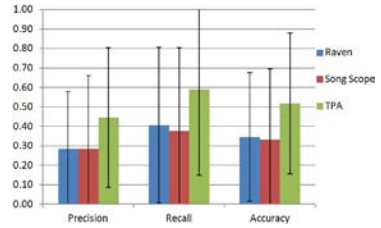


Figure 7. Comparison among Raven, Song Scope and TPA.

Clearly, TPA outperforms Raven and Song Scope under the same conditions during dawn chorus (4am to 9am). Precision, recall, and accuracy all have dramatic increase. The error bars show the distribution of precision, recall and accuracy. As we can see, the distribution is consistent among three indices. However, the one standard deviation is high. This is because the testing data is from dawn chorus when many species

call at once. In total of 114 minutes where there were whip bird calls, the signal is either too weak or too noisy; this noise causes tools fail to detect. Therefore, the precision and recall are all zero. Zero precisions have a strong negative impact on the mean and increase the standard deviation. Even the rest of minutes have better recognition results, the percentage over the total minutes is small. Table IV shows the number and percentage of zero precisions after detection of tools. From this table, we are convinced that even under noisy situation, the recognition ability of TPA is still better than Raven and Song Scope. However, we admit that detecting targets during dawn chorus is really a difficult research problem: the accuracy of TPA is still only 0.52.

Table IV. The number of zero precisions

	Raven	Song Scope	TPA
Precision (0)	44	58	32
Percentage Over total minutes (114)	39%	51%	28%

6 Discussion and Conclusion

Raven and Song Scope are well-developed tools for terrestrial animal detections. Through running the recordings collected in the real environment on them, we surveyed their operation procedures and analyzed the statistical results.

In theory, Raven explores two different detectors to locate the syllables in the spectrogram while Song Scope can detect the call structures using feature vector and HMM. Second, compared to Song Scope, Raven has a friendly interface and more powerful control modules. Because Song Scope requires expertise about signal processing to configure parameters, this makes it more difficult to use than Raven. Significantly, in terms of the recognition ability for five types of call components, Raven has relatively better performance than Song Scope with accuracy of 0.43 and 0.37, respectively. The precision of Song Scope is higher but the recall is lower. This indicates that Raven can be applied to detect the activity of animals while Song Scope to detect the presence of a target.

Instead of detecting syllables only on Raven and just call structures on Song Scope, TPA is designed not only building acoustic component (syllable) detectors separately, but also using Syntactic Pattern Recognition and Markov chains to cluster the components in order to form call structures, which can provides users the initiative to run the component filters and build call structures according to their specific targets by themselves. Compared to Song Scope and Raven, the precision, recall and accuracy are all dramatically increased with TPA. Even in the noisy environment (dawn chorus), TPA picks up extra 10% signals than them.

This paper is part of an ongoing research project for automatic species recognition. The TPA approach is still under testing and construction. More experiments need to be involved such as dataset from multiple sites and multiple days, more species with complex call structures. Actually, it is a difficult task to recognize targets during dawn chorus in automated species call recognition research area based on existing machine learning techniques. Even TPA has approximately 50% accuracy. It has not yet reached a level of reliability that allows ecologists to use the methods without careful verification of results. Much work is required for the real applications in future.

References

Agranat, I. (2009). Automatically Identifying Animal Species from their Vocalizations. Paper presented at the Fifth International Conference on Bio-Acoustics.

Berwick, R. C., Okanoya, K., Beckers, G. J. L., & Bolhuis, J. J. (2011). Songs to syntax: the linguistics of birdsong. *Trends in Cognitive Sciences, 15*(3), 113-121.

Bioacoustics Research Program, 2011. Raven Pro: Interactive Sound Analysis Software (Version 1.4) [Computer software]. Ithaca, NY: The Cornell Lab of Ornithology. Available from http://www.birds.cornell.edu/raven.

Charif, RA, LM Strickman, AM Waack., 2010. Raven Pro 1.4 User's Manual. The Cornell Lab of Ornithology, Ithaca, NY.

Crothers, L., Gering, E., & Cummings, M. (2011). Aposematic Signal Variation Predicts Male–Male Interactions in A Polymorphic Poison Frog. *Evolution*, 65(2), 599-605.

Depraetere, M., Pavoine, S., Jiguet, F., Gasc, A., Duvail, S., & Sueur, J. (2012). Monitoring animal diversity using acoustic indices: Implementation in a temperate woodland. Ecological Indicators, 13(1), 46-54.

Duan, S., Towsey, M., Zhang, J., Truskinger, A., Wimmer, J., & Roe, P. (2011, 6-9 Dec. 2011). Acoustic component detection for automatic species recognition in environmental monitoring. Paper presented at the Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP), 2011 Seventh International Conference on.

McCallum, A. (2010). Birding by ear, visually. Part 1: Birding acoustics. *Birding*, 42, 50-63.

Olson, D. L., & Delen, D. (2008). Advanced Data Mining Techniques (1 ed.): Springer.

Scott Brandes, T. (2008). Automated sound recording and analysis techniques for bird surveys and conservation. *Bird Conservation International*, 18(S1), S163-S173.

Somervuo, P., Harma, A., & Fagerlund, S. (2006). Parametric Representations of Bird Sounds for Automatic Species Recognition. IEEE Transactions on Audio, Speech, and Language Processing, 14(6), 2252-2263.

Song Scope 4.0 User's Manual, 2011. Wildlife Acoustics, Inc, USA.

Stowell, D., & Plumbley, M. (2011). Birdsong and C4DM: A survey of UK birdsong and machine recognition for music researchers: Centre for Digital Music, Queen Mary, University of London.

Towsey, M., Planitz, B., Nantes, A., Wimmer, J., & Roe, P. (2012). A toolbox for animal call recognition. *Bioacoustics*, 1-19.

Wildlife Acoustics, 2011. Song Scope: Bioacoustics Software (Version 4.1.1) [Computer Software]. USA: Wildlife Acoustics, Inc. Available from http://www.wildlifeacoustics.com/products/analysis-software.

Zhuang, X., Zhou, X., Hasegawa-Johnson, M. A., & Huang, T. S. (2010). Real-world acoustic event detection. *Pattern Recognition Letters*, *31*(12), 1543-1551.